Title
Systems-level analyses identify extensive coupling among gene expression machines.

Journal
Molecular systems biology, 2(1)

ISSN
1744-4292

Authors
Maciag, Karolina
Altschuler, Steven J
Slack, Michael D
et al.

Publication Date
2006

DOI
10.1038/msb4100045

Peer reviewed

Molecular Systems Biology (2006) doi:10.1038/msb4100045

www.molecularsystemsbiology.com
Article number: 2006.0003

molecular systems biology

Systems-level analyses identify extensive coupling among gene expression machines

Karolina Maciag[1], Steven J Altschuler[2], Michael D Slack[2], Nevan J Krogan[3], Andrew Emili[3], Jack F Greenblatt[3], Tom Maniatis[4,*] and Lani F Wu[2,*]

[1] Bauer Center for Genomics Research, Harvard University, Cambridge, MA, USA, [2] Department of Pharmacology and Green Comprehensive Center for Molecular, Computational and Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA, [3] Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada and [4] Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA
* Corresponding authors. T Maniatis, Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA. Tel.: +1 617 495 1811; Fax: +1 617 495 3537; E-mail: maniatis@mcb.harvard.edu or LF Wu, Department of Pharmacology and Green Comprehensive Center for Molecular, Computational and Systems Biology, University of Texas Southwestern Medical Center, 6001 Forest Park Blvd, Mail code 9041, Dallas, TX 75390, USA. Tel.: +1 214 645 6182; Fax: +1 214 645 5982; E-mail: lani.wu@utsouthwestern.edu

Received 14.6.05; accepted 6.12.05

Here, we develop computational methods to assess and consolidate large, diverse protein interaction data sets, with the objective of identifying proteins involved in the coupling of multicomponent complexes within the yeast gene expression pathway. From among ~43 000 total interactions and 2100 proteins, our methods identify known structural complexes, such as the spliceosome and SAGA, and functional modules, such as the DEAD-box helicases, within the interaction network of proteins involved in gene expression. Our process identifies and ranks instances of three distinct, biologically motivated motifs, or patterns of coupling among distinct machineries involved in different subprocesses of gene expression. Our results confirm known coupling among transcription, RNA processing, and export, and predict further coupling with translation and nonsense-mediated decay. We systematically corroborate our analysis with two independent, comprehensive experimental data sets. The methods presented here may be generalized to other biological processes and organisms to generate principled, systems-level network models that provide experimentally testable hypotheses for coupling among biological machines.
Molecular Systems Biology 17 January 2006; doi:10.1038/msb4100045
Subject Categories: computational methods; chromatin & transcription
Keywords: protein interactions; network analysis; gene expression; transcription; RNA processing

Introduction

Gene expression is a stepwise process involving distinct cellular complexes that carry out each subprocess, including transcription, pre-mRNA capping, splicing and polyadenylation, mRNA export, and translation. Recent studies have revealed both physical and functional interactions between many proteins involved in separate subprocesses (Maniatis and Reed, 2002; Orphanides and Reinberg, 2002; *Curr Opin Cell Biol* 2005**;** **7**: 239–339). This coupling is thought to improve quality, efficiency, and timing. The tight coupling of gene expression machines can also facilitate rapid changes in gene expression patterns (Misteli, 2001). For example, the carboxy-terminal domain of RNA polymerase II is able to sequentially associate with diverse components of mRNA processing machinery, thus localizing key components to the nascent pre-mRNA in a coordinated fashion (Proudfoot *et al*, 2002; Bentley, 2005).

Although the experimental evidence for the coupling of gene expression subprocesses is mounting, systematic searches for individual components that facilitate this coupling have just begun to be conducted (Burckin *et al*, 2005). Recent advances in high-throughput experimental technologies have led to the creation of genomic-scale protein–protein interaction data sets, which carry with them the possibility of discovering many new functional coupling relationships. However, these data sets also bring challenges due to their large size, high false-positive rates of reported interactions (Edwards *et al*, 2002), and experimental biases leading to variable coverage of the proteome. One approach to ameliorate the biases and gaps inherent to each source is to combine multiple data sets, allowing a survey of experimental techniques and conditions (Bader and Hogue, 2002). However, this approach risks masking true coupling connections with the accumulation of noise from false positives (Gerstein *et al*, 2002). As with any automated analysis of large data sets, the challenge lies in discerning prioritized and biologically interpretable results.

Here, we present a multistep approach to identify coupling between gene expression subprocesses, designed to overcome the above challenges. We needed to (1) select, assess, and combine data sets, (2) cluster the interaction network to suggest protein groupings, and (3) find motifs that indicate coupling among the clusters. To avoid skewing disparate steps in analysis to observations of the final results, we did not choose parameters based on the overall final result. Rather, our

guiding principle was to select parameters at intermediate steps to satisfy appropriate, objective criteria. For example, criteria for data integration include minimizing the impact of adding data sets with high false-positive rates to the combined data set; criteria for computing protein clusters include maximizing functional consistency.

For step 1, we focused our analysis on yeast to take advantage of the large amount of available yeast protein interaction data and of the high degree of conservation of yeast gene expression mechanisms to their well-studied mammalian counterparts. The construction of interactome networks from diverse data sources is an important area of active research (Grigoriev, 2003; Lee *et al*, 2004; Li *et al*, 2004). Our collection of 13 protein interaction data sets (Table I) aims for as complete a survey of publicly available interaction data as reasonably possible. First, we included the results of high-throughput, whole-proteome screens (Table I, data set IDs 1–4); the currently available data sets use the yeast two-hybrid and protein complex immunoprecipitation techniques. We supplement these with data sets of computationally predicted protein interactions (Table I, data set IDs 5–7). Finally, we included results from traditional, hypothesis-driven experiments described in the literature (Table I, data set IDs 8–13). These last sets were collected from the interaction repository database at the Munich Information Center for Protein Sequences (MIPS) (Mewes *et al*, 2002) with high-throughput experiments (1–4) removed. Proteins not linked to proteins functionally annotated to gene expression were excluded from the final data set (Supplementary information 1 and 2).

Current methods for relative data set reliability assessment include comparisons to 'gold standards' (Jansen *et al*, 2003), expression profiles, and functional conservation (Deane *et al*, 2002). Multiple data sets can be integrated in several ways, such as the Bayesian (Jansen *et al*, 2003) and random forest (Qi *et al*, 2005) methods used in other studies. Because we wanted to be able to flexibly adapt to growing numbers of data sets without reliance on a gold standard, we developed a relative method for determining data set reliability based solely on mutual comparison. We then cumulated evidence for each individual network edge/protein pair interaction in proportion to the computed reliability of its source.

Table I List of protein interaction data sets used

Data set ID	Data set name	No. of distinct proteins	No. of distinct links	Computed RDQ score
1	Ito-core Y2H	225	1675	5.398
2	Ito-full Y2H	721	400	1.160
3	Uetz Y2H	282	462	6.842
4	Complex	763	4753	2.472
5	Rosetta	161	905	0.262
6	Paralog	1508	32 017	0.581
7	Phylogenetic	21	71	1.477
8	MIPS-affinity	62	79	28.661
9	MIPS-co-precipitation	187	385	17.318
10	MIPS-co-purification	98	208	18.920
11	MIPS-synthetic	341	681	5.919
12	MIPS-Y2H	425	961	8.793
13	MIPS-other	532	2912	2.196

For each of the 13 data sets, we show the number of distinct proteins (column 3) and pairwise protein interactions (column 4). RDQ scores are computed as described in the text. The larger the RDQ score of a data set, the greater its contribution to the combined interaction network.

For step 2, multiple methods have been developed to identify 'modules' in protein interaction networks, including clustering the interaction network alone (Samanta and Liang, 2003; Spirin and Mirny, 2003; Tornow and Mewes, 2003), or introducing orthogonal classes of data (Yeger-Lotem and Margalit, 2003; Lee *et al*, 2004; Tanay *et al*, 2004; Gunsalus *et al*, 2005). We chose to use the unsupervised *k*-means algorithm, which is intuitive to understand and interpret, and is commonly used to cluster continuous-valued data, and modified it to work on network data. Finally, for step 3, biologically significant motifs within networks have been defined by methods such as the least or most connected subgraphs (Vazquez *et al*, 2004), or discovered using metrics such as recurrence frequency (Shen-Orr *et al*, 2002). Motivated by these approaches, we developed a method to identify motifs of coupling among clusters, rather than among individual proteins. Our approach used a supervised method to find instances of motifs modeled on known patterns of biological coupling, rather than empirically discovering *de novo* coupling motifs.

Results

Summary of method

We began by consolidating diverse binary protein interaction data sets into a single network, where nodes represent proteins and edge weights reflect the accumulation of evidence supporting their interaction (Figure 1A). First, each of the 13 data sets was formatted, if necessary, to indicate only the presence or absence of a pairwise protein interaction. Each data set was then reduced to include only interactions among the 2100 proteins that may be involved in gene expression, based on annotations in Gene Ontology and augmented by MIPS data (Figure 1A(1); Supplementary information 1 and 2; Supplementary Table S1). We then assessed the reliability of each data set to compute a relative data set quality (RDQ) score, based on pairwise comparisons of its mutual overlap with every other data set (Figure 1A(2); Supplementary information 5; Supplementary Table S2). This concept of evaluating quality based solely on mutual comparison has been used by search engines for ranking web pages (Page *et al*, 1998). Notably, as this method relies on checks and balances among data sets, RDQ evaluations perform better when larger numbers of data sets with diverse biases are compared. Each data set contributes to the final, integrated network, by adding links appropriately weighted by the corresponding RDQ (Figure 1A(3) and Supplementary information 3).

To reduce the weights of links due to false positives, and to reconstruct links missed due to false negatives, we generalized the pairwise clustering coefficient (CC) (Goldberg and Roth, 2003) to weighted networks. Our CC formula is a measure of the local, weighted network neighborhood around a pair of proteins, including pairs lacking a direct link (Figure 1A(4) and Supplementary information 8). Heuristically, for each pair of proteins, links to common neighbors increase the CC, whereas

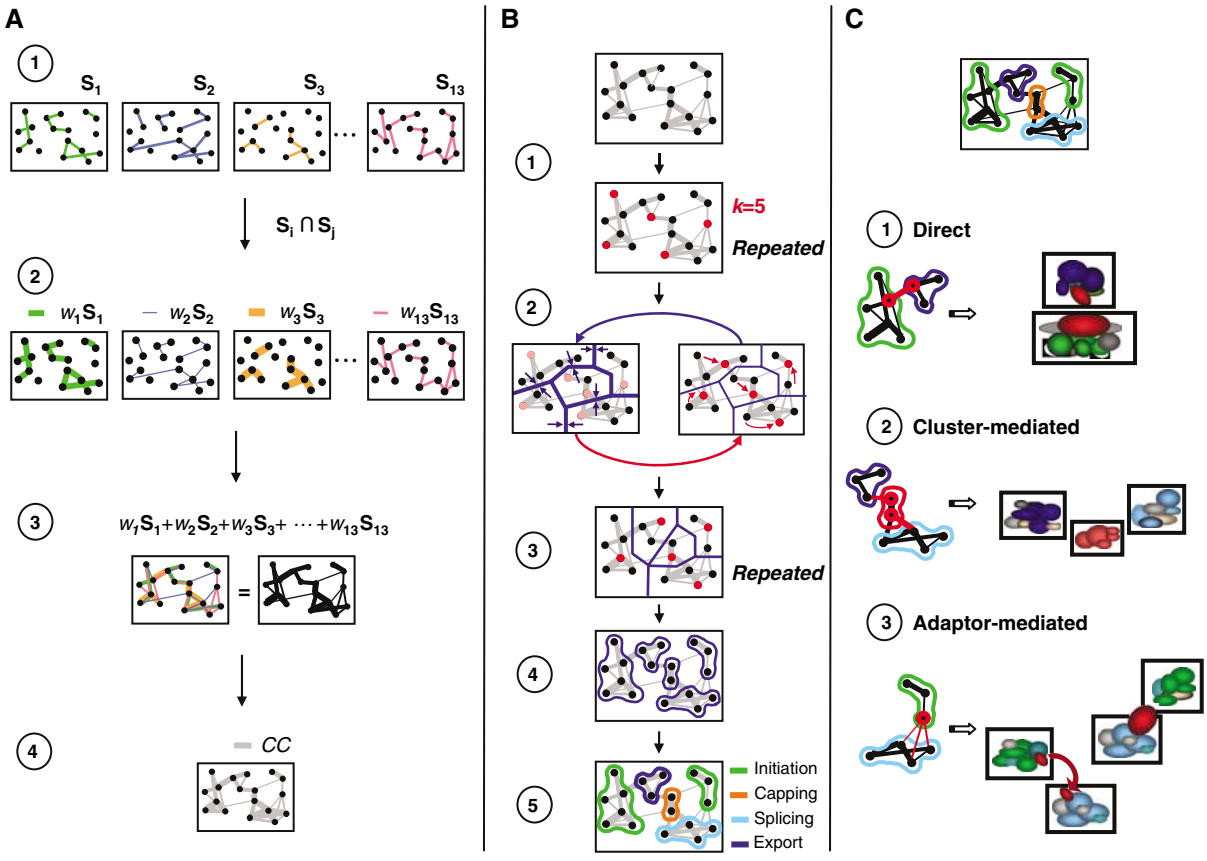

Figure 1 Overview of method (see main text and Supplementary information for details). (**A**) Construction of an integrated protein interaction network. Nodes represent proteins and links represent protein interactions. Line thickness corresponds to link weight (*w*). Input (1), relative quality calculation (2), and integration (3) of networks defined by interaction data sets S1–S13 generate a comprehensive, weighted protein interaction network. Pairwise CC scores (CC) are computed (4) using local network weight and topology information. (**B**) Unsupervised identification of biologically significant clusters in the network using an iterative clustering algorithm based on CC scores. Each randomized selection of *k* initial centers (1, in our analysis, k=70) followed by iterations of cluster definition and center repositioning (2) yields a clustering (3); the best clustering from multiple trials, as defined in the text, is chosen (4). Clusters generated are functionally characterized (5). (**C**) Motifs in the interaction network identify direct (1), cluster-mediated (2), and adaptor-mediated (3) coupling among clusters.

links to uncommon neighbors indicate promiscuous binding and decrease the CC.

To find molecular machineries corresponding to clusters of densely connected proteins in our weighted network, we developed a modification of the *k*-means clustering algorithm (Figure 1B). First, *k* initial 'centroid' proteins are randomly chosen, and each remaining protein is assigned to the centroid to which it is most strongly linked based on cumulative path weight, defined as the sum of CC-weighted links along the shortest path to the centroid. Each centroid thus defines a discrete cluster of associated proteins (Figure 1B(1) and 1B(2), right), and Supplementary information 10). We then (1) reassign the centroid of each cluster to the protein of maximal average CC to all other proteins within the cluster (hence, 'centroid') (Figure 1B(2), right), (2) reassign proteins into clusters about the new set of centroids (Figure 1B(2), left), and (3) iterate (1, 2) until convergence (Figure 1B(3)). We repeated this procedure with many sets of randomly chosen initial centroids and chose the clustering that maximized CC-weighted paths to centroids for subsequent steps (Figure 1B(4)). We then annotated this final set of clusters based on functional enrichment in gene expression subprocesses (Figure 1B(5) and Supplementary information 12) (Tavazoie *et al*, 1999; Wu *et al*, 2002).

Finally, we searched the network for three types of motifs, or patterns in the arrangement of links and clusters that reflect observed manners of coupling between protein complexes (Figure 1C). Direct coupling identifies strong links between proteins tightly grouped within separate clusters, such as the interactions between subunits of the Mediator complex and the basal transcriptional machinery (Kuras *et al*, 2003) (Figure 1C(1)). Cluster-mediated coupling identifies small clusters that link two larger clusters, such as the exon junction complex coupling splicing and export machineries in mammals (Reed, 2003) (Figure 1C(2)). Adaptor-mediated coupling identifies proteins that may belong to either of two clusters, such as scaffolding linker proteins or proteins that shuttle between complexes and transiently associate with each (Figure 1C(3)). For example, the yeast transcription termination and poly-adenylation machineries are thought to be linked by the protein Ctk1p, which is involved in both processes (Kim *et al*, 2004). Our method ranked all instances of these motifs in the network to present a prioritized list of experimentally verifiable biological hypotheses.

A novel algorithm for protein data set reliability calculation

Current approaches to assessing the quality of high-throughput data sets generally determine the rate of false positives and/or false negatives of each test data set with respect to a benchmark 'gold standard' (Jansen and Gerstein, 2004). A typical benchmark has been the MIPS data set, a repository for protein interaction data from the literature (Mewes *et al*, 2002). Our challenge was to develop a method for assessing RDQ scores of our data sets independent of such a gold standard. Intuitively, an RDQ score for a data set can be determined as a function of how well it corroborates with each of the other data sets, modulated by the RDQ scores of each of the corroborating data sets. Thus, RDQ scores can be computed by starting with an initial vector of RDQ scores, iteratively corroborating data sets with each other, and dynamically updating their RDQ scores until the scores converged. In fact, given a matrix M reflecting the pairwise corroboration of the data sets, this converged RDQ weight vector is given by the principal eigenvector of M. Given a set of RDQ scores w_i, reflecting the relative trust of each data source with respect to the others, the individual data sets of protein interactions S_i (represented by matrices with binary entries) can be combined by simple matrix addition to form a final graph $S=\sum w_i S_i$ (Figure 1A).

We compared RDQ scores generated using three different measures of data set overlap (Supplementary information 5) by their ability to down-weight data sets with noise due to false positives (Supplementary Figure S1 and Supplementary information 6), a problem pervasive in some genomic-scale data sets (von Mering *et al*, 2002). The set of RDQ scores chosen according to this criterion also rewards data sets enriched in links between functionally related proteins (Supplementary Figure S2 and Supplementary information 7). Based on these scores, the MIPS affinity and MIPS co-purification data sets were shown to be the most reliable, whereas the Rosetta fusion and paralog-predicted data sets were the least reliable, as might be expected (Table I).

Clustering coefficient

As discussed above, the pairwise CC measures the local neighborhood cohesiveness around a pair of proteins by assessing relative connectivity to common and distinct neighbors, allowing us to account for false negatives and compensate for false positives (Goldberg and Roth, 2003). Our challenge was to generalize the concept of CC to weighted graphs to incorporate link weight as well as topology. We developed and compared a range of CC formulas and selected the formula with scores that reflect the strongest correlation to functional relationships between proteins. By this measure, the selected CC scores significantly outperform even the original network weights (Supplementary Figure S3; Supplementary Table S3; Supplementary information 8 and 9).

Biological interpretation of clusters generated

We used a modification of the k-means deterministic clustering algorithm (Hartigan, 1975) to partition our network of 2100 proteins (Supplementary information 10). This algorithm produces k discrete clusters of proteins linked by strong CC weights and is well suited to the identification of protein complexes corresponding to hypothesized molecular machines. We surveyed different numbers of desired final clusters, k, and found that k=70 consistently generated biologically interpretable clusters (Supplementary information 11), with an average of 30 proteins per cluster. The resulting 70 clusters used for further analysis (Figure 2A(ii) and Supplementary Table S5) were annotated based on enrichment in our nine defined subprocesses of gene expression (Figure 2A(iii) and Supplementary Table S6). Out of the 70 clusters, 48 showed significant enrichment in one or more subprocesses of gene expression (P-value <0.05; see Materials and methods), and all nine subprocesses were represented by at least one cluster (Figure 2C). Among these 70 clusters, ~25% were densely interconnected as expected. However, ~50% of the clusters were sparsely linked internally but contained proteins linked through a small number of common binding partners (some of which appear in different clusters). This was due to the use of pairwise CC scores in the clustering process, which assigns high scores to pairs of proteins having common neighbors regardless of the existence of a direct link between them.

Dense internal linkage

Many of the densely internally linked clusters were found to correspond to well-characterized biological complexes in yeast (Figure 3A), the members of which are highly similar to their counterparts in mammalian complexes. For example, 64% of the known yeast spliceosomal proteins were automatically grouped together in cluster #21 (abbreviated as C21). Other co-clustered complex members include the mRNA cleavage/polyadenylation factors (C26), subunits of the CCR–NOT complex (C27), and the chromatin remodeling machineries SAGA, Swi/Snf, ISWI, and RSC (C1, Figure 3B, left).

Sparse internal linkage

Many of the more sparsely internally linked clusters reconstruct known functional modules that may represent sets of conditionally associated or interchangeable parts that bind to common partners. For example, our clustering grouped the general transcription factor (GTF) TFIIA in cluster C8 together with subunits of the GTF TFIID despite lack of interaction in the original data sets (Figure 3B). Indeed, the interaction between one of these TFIID subunits with TFIIA has just recently been validated experimentally (Robinson *et al*, 2005). Interestingly, the subunits of TFIID are all either in the sparsely interconnected cluster C8 or in the densely interconnected cluster C1. The interconnected subunits in C1 represent a TFIID core, which includes the yeast TBP homolog Spt15p, subunits common to TFIID and SAGA (TAFs), and two other subunits central to the TFIID assembly (Taf1p/Taf2p) (Auty *et al*, 2004). Spt15p is shown to bind each TFIID subunit by an interaction in at least one of our data sets, suggesting that the mutually unconnected subunits in C8 may conditionally associate with the TFIID core in C1. This is consistent with

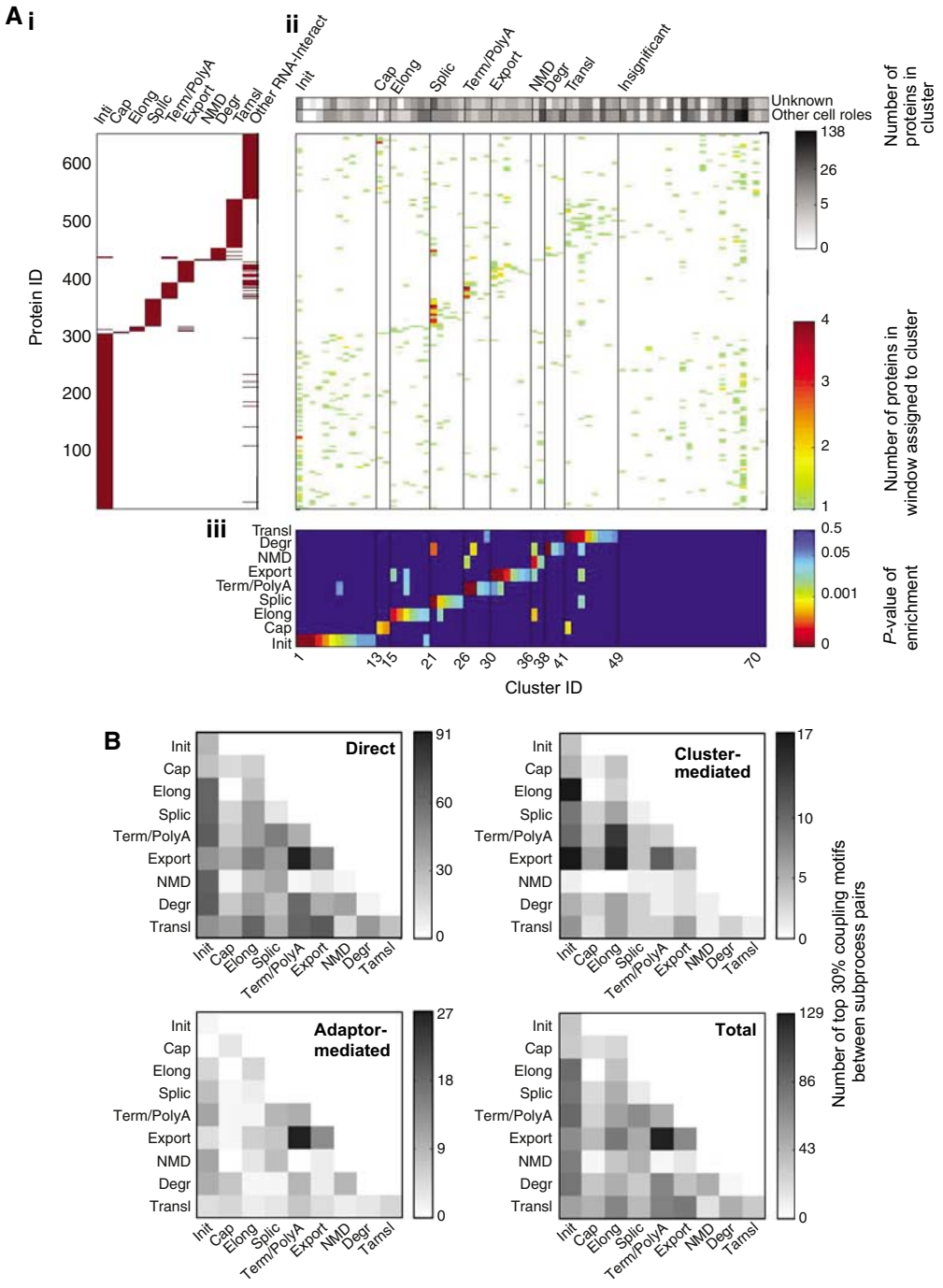

Figure 2 (**A**) Relationship between proteins, clusters, and annotation to gene expression subprocesses. (i) Assignment of proteins (vertical axis) to subprocesses (horizontal axis, labeled). Note that each protein may be annotated to more than one subprocess. Proteins are ordered along the vertical axis by the least abundant subprocess to which they belong. (ii) Assignment of proteins to clusters (horizontal axis). Proteins (vertical axis, as in (i)) are grouped into segments along the vertical axis containing four proteins at a time. For each segment, the number of proteins (out of a possible four) assigned to each cluster is shown, as indicated by the color bar. The frequencies of proteins in the cluster annotated to other, non-gene expression cell roles and proteins of unknown function are indicated in grayscale (top). Clusters are ordered along the horizontal axis by predominant functional annotation as determined in (iii). (iii) Annotation of clusters (horizontal axis) to subprocesses (vertical axis). The plot shows the *P*-value of statistical significance of enrichment of each cluster in proteins annotated to each subprocess. Clusters are ordered along the horizontal axis by predominant subprocess annotation. (**B**) Distribution of coupling among pairs of gene expression subprocesses. For each pair of subprocesses, plots indicate the frequency of cluster pairs that are significantly annotated to the respective subprocesses ($P < 0.05$) and are linked by coupling motifs ranked in the top 30% of direct (top left), cluster-mediated (top right), and adaptor-mediated (bottom left) motifs. Frequencies of motif instances are indicated in grayscale. The sums of links using all three motifs are shown as well (bottom right).

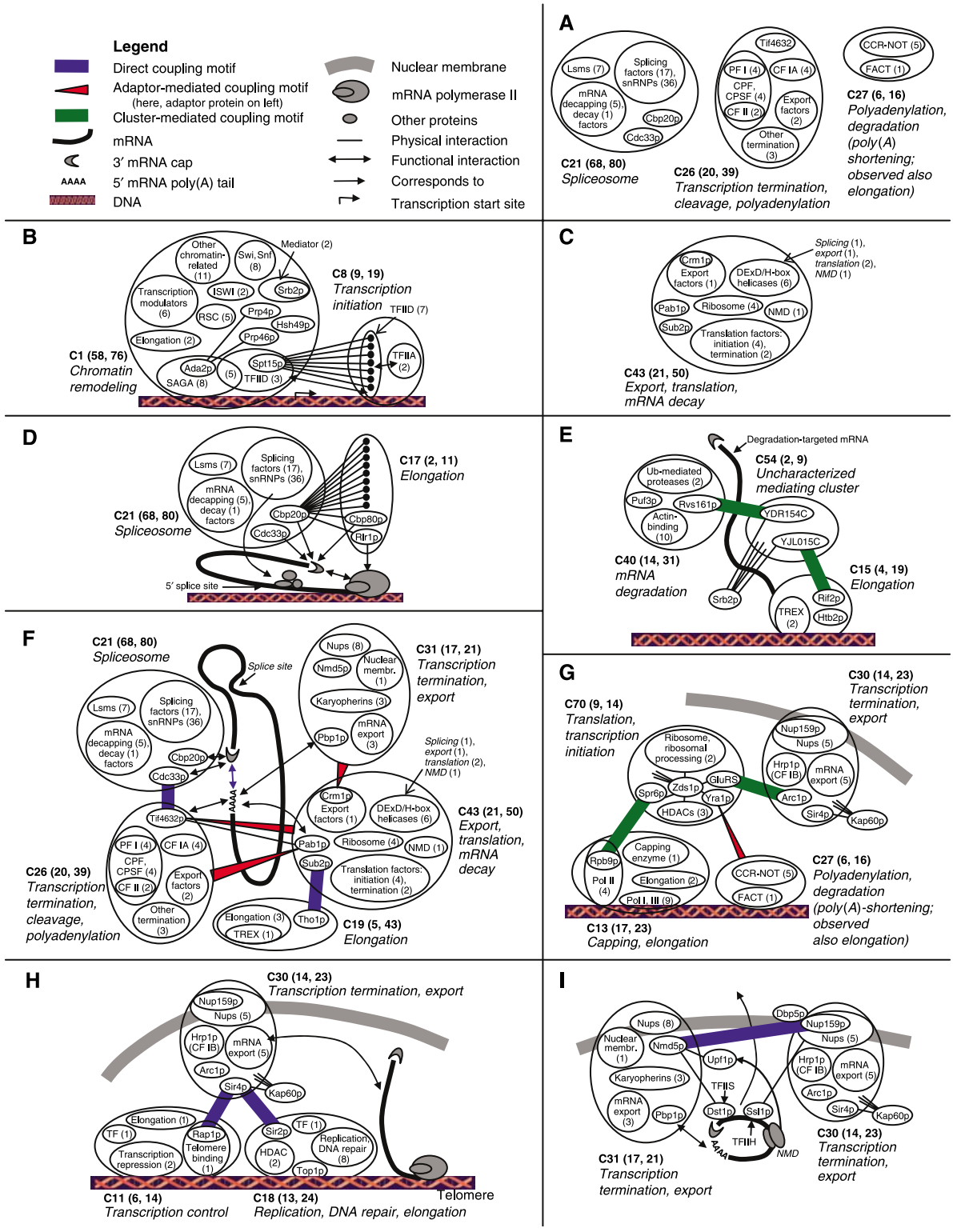

Figure 3 Protein clusters and top-ranking motifs suggest mechanisms of coupling between gene expression processes. Motifs described in the text are illustrated. For each cluster, *n* indicates the total number of proteins illustrated and *m* the total number of proteins in the cluster. For each specially noted subgroup of proteins within a cluster, *P* indicates the number of proteins in the subgroup. (**A**) Clusters may reconstruct well-known structural complexes. (**B**) Clusters reconstruct GTF machinery despite data missing from original data sets, and suggest conditional association of members in C8. (**C**) Seven DExD/H helicases in a single cluster identify a functional module. (**D**) Coupling of capping, elongation, and splicing suggested by the co-clustering and binding patterns of a cap-binding protein. (**E**) Top-ranked cluster-mediated coupling motif suggests coupling between elongation and mRNA quality control degradation. (**F**) Direct and adaptor-mediated coupling motifs suggest possible nuclear mRNA circularization, along with coupling among mRNA transcription and processing with export. (**G**) Cluster- and adaptor-mediated coupling motifs suggest coordination of transcription, export, and translation. (**H**) The top-ranking direct coupling motifs indicate possible coupling of mRNA export to chromatin silencing. (**I**) Direct coupling motif and co-clustering suggest coupling of transcription and mRNA export with translation and NMD, possibly at the nuclear pore.

the fact that Spt15p is sufficient for basal transcription (Kim *et al*, 1994), whereas different TFIID subunits are required for transcription of distinct gene subsets (Walker *et al*, 1997). Similarly, use of the CC co-clustered the seven DExD/H-box helicases (Figure 3C), which independently carry out similar biological roles (oligoribonucleotide unwinding) in nearly every step of mRNA processing (de la Cruz *et al*, 1999), in the sparsely interconnected cluster C43.

Co-clustering due to direct physical interaction or links to common external partners is consistent with subprocess coupling. The co-clustering of the cap-binding protein Cbp20p with splicing factors in cluster C21 due to physical interactions, for example, suggests coupling of capping and splicing (Figure 3D). Cbp20p is a common binding partner for co-clustering of the elongation protein Rlr1p, the cap-binding protein Cbp80p, and other proteins in the sparsely interlinked cluster C17, suggesting the coupling of capping and elongation. Indeed, Cbp20p and Cbp80p have been shown to be involved in splicing experimentally (Colot *et al*, 1996; Lewis *et al*, 1996; Hirose and Manley, 2000). Together, these results suggest that Cbp20p and Cbp80p (along with Cbc33p, another cap-binding protein in C21), bound to the nascent 5′ cap, are associated with elongation complexes and poised to interact with splicing components. This is consistent with previous studies showing coupling among elongation, capping, and splicing machineries (Maniatis and Reed, 2002; Orphanides and Reinberg, 2002).

Intercluster interaction motifs

To identify potential coupled protein machineries, we identified cluster pairs or triples with interactions that satisfy motifs identified as hallmarks of process coupling. To measure the degree to which cluster pairs represent distinct (yet potentially coupled) protein complexes, rather than single molecular machineries artificially separated by our clustering process, we developed a pairwise cluster separability score (Supplementary information 13). For a pair of clusters, the cluster separability score is simply the sum of the individual *k*-means clustering scores divided by the clustering score of their merger; the bigger the ratio, the more 'natural' the separation of the clusters. We imposed a threshold so that only cluster pairs with separability score in the top 50% (thus more likely to represent distinct machineries) were searched for coupling motifs. From these, we identified and ranked 2029, 517, and 276 direct, cluster-mediated, and adaptor-mediated coupling motifs, respectively (Figure 1C and Supplementary Tables S7–S9), for further biological investigation. Rankings were based on network link strength and topology to prioritize motifs by potential biological relevance. Top 25-ranked coupling motifs and cluster protein compositions are visualized in Supplementary Figure S4. The robustness of identification and ranking of motifs to the selection of an optimum clustering run is evaluated in Supplementary information 14.

Experimental corroboration

Previous studies have led to the establishment of yeast complex precipitation data sets using tandem-affinity purification followed by mass spectrometry (Gavin *et al*, 2002; Ho *et al*, 2002). From these, we inferred pairwise protein interactions as described in Supplementary information 2. New, more comprehensive genome-wide complex precipitation data sets were recently obtained in which proteins were identified using gel-free liquid chromatography mass spectrometry (LCMS) or matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (Krogan *et al*, in preparation). A database of pairwise protein interactions was derived for each method in which the interactions were quantified by confidence or reliability scores. Because these data sets were not used to generate our model of coupled protein clusters, they provide an opportunity to corroborate our analysis with independent, experimental protein interaction data.

For each of the two independent data sets, this corroboration can be illustrated by the fold enrichment: the ratio of the number of interactions in the data set that define direct coupling motifs to the number of these interactions defining direct coupling motifs in a randomized model (Supplementary information 15). For a complete analysis, we carried out this corroboration after applying four different thresholds on confidence scores in the independent data sets (Figure 4). More stringent thresholds, which presumably select higher quality data, provide a better fit to our model by demonstrating greater fold enrichments. Fold enrichment was improved by considering successively higher ranked coupling links in our model. Using the most stringent threshold, enrichments as high as 350-fold were observed for the top 1% of direct coupling links; the top 30%, however, still showed a $\sim$50-fold enrichment. The peak around 15–25% is likely due to the fact that many direct interactions ranked in the top 8–13% have equal ranking score and are ranked arbitrarily; of these, many may be matched by false negatives in the corroborating data sets, especially the LCMS set.

Similar corroboration was performed for interactions defining cluster-mediated and adaptor-mediated motifs, and significant, although lower, fold enrichments were obtained (Supplementary Figure S5). Furthermore, interactions from the independent data sets with highest stringency thresholds applied were approximately seven times more likely to fall within clusters, two times more likely to occur between coupled clusters, and three times more likely to occur between non-coupled, co-annotated clusters in our model compared to randomized models. Interestingly, comparable fold enrichments were found using either the LCMS or the MALDI data sets. Therefore, independent data quantitatively support the predictive power of our approach, both in the formation of coupled, functionally annotated clusters and especially in the identification and ranking of coupling links between them.

Top-ranking results confirm and infer coupling among gene expression machineries

The distribution of motifs ranked in the top 30% of their respective motif pattern among pairs of gene expression subprocesses is shown in Figure 2B. In the paragraphs below, we assess the biological significance of coupling interactions indicated by top-ranked motifs with existing literature and suggest potential new coupling mechanisms in the contexts of chromatin dynamics, quality control, and mRNA export.

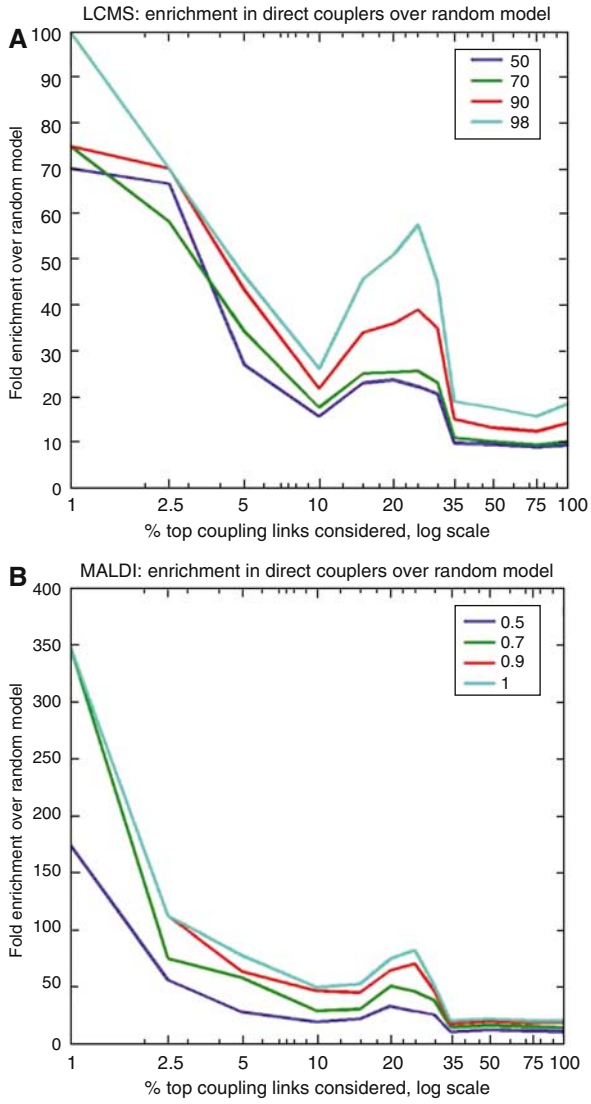

Figure 4 Fold enrichment of interactions in independent protein interaction data sets identified as direct coupling links in our model, as compared to randomized models. The independent, comprehensive protein interaction data sets were derived from systematic, previously unpublished complex precipitation studies using (**A**) LCMS and (**B**) MALDI-TOF mass spectrometry analysis. Shown are the fold enrichments of the number of interactions identified as direct couplers in the model used in this analysis, over the average number of interactions identified as direct couplers in 50 randomized models. The fold enrichment (*y*-axis) is shown as a function of the percentage of top-ranking direct coupling links considered (*x*-axis). Higher ranking links are more likely to appear in the independent data sets. Independent protein interaction data sets are subjected to thresholds at four different interaction confidence values (line colors). Higher quality interaction data demonstrates greater enrichment in our model versus in random models.

Pre-mRNA quality control via coupling of mRNA degradation to transcription

While accurately transcribed and processed mRNAs are targeted for export, abnormal pre-mRNAs are degraded due to transcriptional quality control mechanisms. Studies suggest that quality control machinery may associate with nascent pre-mRNAs as soon as transcriptional elongation begins, and travel with RNA polymerase to the termination and polyadenylation site (Minvielle-Sebastia and Keller, 1999). The TREX (transcription and export) complex could provide a potential scaffold to facilitate this surveillance. Notably, the TREX complex also provides a specific example of differences in coupling mechanisms between yeast and mammals (Masuda *et al*, 2005), as its recruitment to nascent mRNA is coupled to transcription in yeast and to splicing in mammals.

The top-ranking cluster-mediated coupling motif implicates C54 as a mediator between the mRNA degradation cluster C40 and the elongation cluster C15, which includes members of TREX (Figure 3E). In particular, the unknown essential protein YJL015Cp, linked to the elongation cluster C15, and the unknown protein YDR154Cp, linked to the mRNA degradation cluster C40, are candidates for involvement in quality control. As all the proteins in sparsely interlinked cluster C54 interact with the transcription initiation protein Srb2p as a common binding partner, this suggests their potential recruitment to sites of transcription initiation. Overall, the existence of this motif suggests that the switch between transcription initiation and elongation is subject to quality control.

Alternate methods of recruitment of the mRNA export machinery to nascent RNA

Nuclear export factors are recruited to nascent pre-mRNA during various steps of RNA processing and transcription (Reed, 2003; Tange *et al*, 2004; Aguilera, 2005; Darzacq *et al*, 2005; Reed and Cheng, 2005). Coupling of export to RNA processing provides a quality control mechanism to prevent the export of incorrectly processed mRNAs (Tange *et al*, 2004). The correct completion of the processing steps of polyadenylation (involved in transcription termination) and splicing, for example, has been shown to be necessary for efficient export in metazoans (Reed, 2003; Sommer and Nehrbass, 2005). Coupling of export to transcription, on the other hand, may poise the machinery for timely and efficient transport. Export is coupled to elongation for pre-mRNAs lacking introns in mammals (Lei *et al*, 2001); this is thought to be a major mechanism in yeast, where over 96% of genes are free of introns. In addition, new results indicate the coupling of export to chromatin modification at the initiation of transcription (Rodriguez-Navarro *et al*, 2004). Our motifs reveal additional support for coupling of export with transcription termination, elongation, and chromatin modification as shown below.

Coupling of mRNA export and transcription termination

First, clusters C31 and C30 both co-cluster termination and export factors, and Hrp1p in C30 is in fact annotated to both processes (Figure 3I). Second, an adaptor-mediated coupling motif (ranked 7th) (Figure 3F, right) links the core termination cluster C26 to the adaptor protein Pab1p, a poly(A)-binding protein in C43, annotated to export, translation, splicing, and mRNA degradation. This coupling role for Pab1p, a member of the termination complex CF IA, is consistent with previous studies in yeast showing that mRNA export is inefficient in CF IA mutants (Hammell *et al*, 2002). Third, another adaptor-coupled motif (ranked 21st) (Figure 3F, top right) links the transcription termination and export module C31 to the adaptor protein Crm1p, an mRNA and protein export factor.

The coupling between transcription termination and mRNA export revealed by our analysis is experimentally supported by the observation that export proteins participate in termination (Jensen *et al*, 2001; Hammell *et al*, 2002).

Coupling of mRNA export and transcription elongation

A direct coupling motif (top 2%-ranked; Figure 3F, bottom) involves the interaction between the mRNA export factor Sub2p (C43) and the putative transcription elongation factor Tho1p in elongation-annotated cluster C19. Also, an adaptor-mediated coupling motif (top 30%-ranked; Figure 3G) links the export protein Yra1p in C70 to C27, containing members of the CCR–NOT elongation complex, thus implicating Yra1p in transcription-coupled export as well. This is consistent with previous observations of elongation-coupled export in which the export proteins Sub2p and Yra1p associate with the THO elongation complex to form TREX, the yeast transcription/export complex (Reichert *et al*, 2002; Strasser *et al*, 2002). In fact, Tho1p is functionally similar to the TREX component Tho2p (Piruat and Aguilera, 1998). Moreover, Sub2p was shown to be essential for export of intronless genes in *Drosophila* (Gatfield *et al*, 2001), and was implicated in coupling elongation and export in a previous genome-wide study (Burckin *et al*, 2005).

Coupling of export and chromatin silencing

The silencing protein Sir4p is assigned to the export-annotated cluster C30 (Figure 3H), based primarily on its physical interaction in the complex data set with Kap60p, an mRNA export factor (Liu *et al*, 1999) that serves as a common binding partner for the members of C30. Top-ranked direct coupling highlights the interactions of Sir4p with two of its known binding partners in clusters enriched in chromatin silencing and DNA-binding proteins: the top direct-coupling motif links Sir4p to the silencing protein Sir2p in C18, and the fourth-ranked direct-coupling motif links Sir4p to the telomere-silencing protein Rap1p in C11. Notably, heterochromatic DNA regions are often found associated with the nuclear periphery in the proximity of nuclear pores (Laroche *et al*, 2000), and Sir2p and Sir4p have previously been localized to the nuclear periphery as well (Huh *et al*, 2003). Although the significance of this localization was previously thought to be a direct functional requisite for gene silencing, recent studies in yeast are inconsistent with this hypothesis (Gartenberg *et al*, 2004). The physical coupling between export and silencing could have other explanations, such as the existence of proteins that participate in both processes, or even localization of silenced regions in proximity to the nuclear pore in order to facilitate export of transcripts from nearby euchromatic regions.

Coupling translation and nonsense-mediated decay with other gene expression processes

Nonsense-mediated decay (NMD) is a translation-dependent quality control mechanism for recognizing and degrading mRNAs containing premature termination codons. NMD in mammalian cells is generally thought to occur in the cytoplasm, although recent studies suggest that it may also occur in the nucleus (Brogna *et al*, 2002; Iborra *et al*, 2004a). Evidence that nuclear translation may occur in yeast exists as well. For example, a potential nuclear pioneer round of translation in yeast is suggested by the binding of the cap-binding complex Cbp20p/Cbp80p to mRNA in the nucleus (Ishigaki *et al*, 2001). In addition, recent studies have shown that the yeast NMD factor Upf1p associates with the nuclear pore, consistent with the possibility that NMD occurs as the mRNA exits the nucleus (Nazarenus *et al*, 2005). However, evidence against nuclear translation in yeast was provided by the observation that blocking export prevents NMD (Kuperwasser *et al*, 2004). While this observation was interpreted in favor of cytoplasm-only NMD, an alternative explanation is that nuclear NMD only occurs when coupled to mRNA export. For example, a 'pioneer round' of translation could occur as the mRNA exits the nuclear pore. Recognizing that nuclear translation in both mammals and yeast is controversial (see Iborra *et al*, 2004b; Dahlberg *et al*, 2004 for contrasting views), we have identified coupling motifs consistent with the possibility that NMD may be coupled to pre-mRNA processing in yeast.

NMD is initiated upon recognition of an untranslated coding mRNA sequence downstream of a termination codon (Hilleren and Parker, 1999; Maquat, 2002). In mammals, a complex of NMD proteins mark coding mRNA sequences by assembling on the exon junctions of newly spliced mRNA to form exon junction complexes (Le Hir *et al*, 2000). By contrast, in yeast, NMD proteins assemble on downstream elements of the coding mRNA (Ruiz-Echevarria *et al*, 1998). Ribosomes participating in a pioneer round of translation in both yeast and mammals are thought to displace the NMD proteins from mRNA. Ribosomes stalled at premature termination codons interact with the NMD complex assembled on downstream mRNA coding sequences, an interaction thought to lead to recruitment of mRNA degradation factors (Czaplinski *et al*, 1998).

Coupling of translation and NMD to transcription and mRNA export

A direct interaction motif (ranked 16th and validated by copurification data; Figure 3I) involves Nmd5p (C31), a protein previously implicated in the NMD process (He and Jacobson, 1995), and the nuclear pore protein Nup159p (C30), consistent with the possibility that NMD occurs at the nuclear pore. Nup159p is known to be involved with the final stage of mRNA export as a docking site for proteins that interact with mRNPs exiting into the cytosol, such as the DEAD-box helicase Dbp5p (Schmitt *et al*, 1999; Weirich *et al*, 2004). Surprisingly, Dbp5p and Nmd5p interact physically with transcriptional elongation factors Dst1p (a TFIIS homolog) (Albertini *et al*, 1998) and Ssl1p (a TFIIH subunit) (Estruch and Cole, 2003), respectively. This suggests co-transcriptional recruitment of factors involved in possible NMD processes at the nuclear pore. Coupling of transcription and NMD is consistent with observations that NMD proteins copurify with several different PolII subunits (NJ Krogan, unpublished data). Nmd5p is known to interact with the key NMD protein Upf1p (Czaplinski *et al*, 1998), and may thus be involved in concentrating Upf1p at sites of pioneer-round translation and NMD. These sites may be nuclear, since in human cells at least, Upf1p has been shown to shuttle to the nucleus (Mendell *et al*, 2002).

Coupling of transcription and translation, and nuclear mRNA circularization

Translation initiation in eukaryotes requires the circularization of mRNA through the binding of cap- and poly(A)-binding proteins with translation initiation factors at the loop junction (Ishigaki *et al*, 2001). The second-ranked direct coupling motif in our analysis suggests that this may occur in the nucleus (Figure 3F). Here, the cap-binding translation initiation factor Cdc33p (C21) binds the poly(A) tail-binding protein Tif4632p (eIF-4F in mammals, C26) (Tarun *et al*, 1997), also involved in translation initiation (Goyer *et al*, 1993; Lang *et al*, 1994). The motif is consistent with the possibility that Cdc33p and Tif4632p associate with the mRNA cap and tail, respectively, then with each other to lead to the circularization of mRNA in the nucleus. Yeast mRNAs associated with either the nuclear Cbp20p/Cbp80p or the cytoplasmic mRNA cap-binding translation initiation factor Cdc33p (eIF-4E) have been shown to be subject to NMD (Gao *et al*, 2005). Yeast Cdc33p has, furthermore, been found to occur in the nucleus as well (Lang *et al*, 1994). Thus, unlike in mammalian cells, both Cbp20p/80p and Cdc33p-bound mRNA may mediate a nuclear pioneer round of translation coupled to NMD in yeast.

Although both proteins were previously characterized as cytoplasmic, Cdc33p and Tif4632p are clustered in predominantly nuclear clusters due to strong physical interactions. This suggests that they could be involved in or recruited through nuclear gene expression events. Experimental results indeed support the presence of Cdc33p in the nucleus in both yeast (Lang *et al*, 1994) and mammals (Wilkinson and Shyu, 2002). The clustering of Cdc33p in the spliceosome cluster C21 is consistent with the possibility that its recruitment to mRNA is coordinated with splicing, as observed for other cap-binding proteins (Lewis *et al*, 1996). Moreover, this motif suggests that the exchange of the (mostly nuclear) Cbp20p/Cbp80p cap-binding translation initiation complex for the (mostly cytoplasmic) Cdc33p occurs in the yeast nucleus in a manner coordinated with splicing and nuclear circularization. This exchange was previously suggested to be nuclear in mammals (Wilkinson and Shyu, 2002). Finally, the poly(A)-binding Tif4632p belongs to the transcription termination cluster C26, perhaps indicating loading onto the nascent poly(A) tail in a manner coordinated with termination.

Coupling of translation and NMD to transcription and mRNA export

The second-ranked cluster-mediated coupling motif involves the mediating cluster C70, identified as a translation module because it is enriched in protein components of the ribosome and ribosome processing factors (Figure 3G). The motif couples C13, a transcription elongation and capping cluster, with C30, an mRNA export cluster. The RNA polymerase subunit Rpb9p in C13 interacts with Spr6p, a protein of unknown function and localization in the mediating cluster, suggesting the participation of Spr6p in coupling transcription to translation in the nucleus. In the mediating cluster, the tRNA synthetase GluRS further links to another protein necessary for translation, the tRNA delivery protein Arc1p, in the export cluster C30. This could be explained in several ways. The first is the possible co-export of NMD proteins with other RNA-associated proteins, leading to a perceived coupling motif. These proteins could also be directly involved in mRNP packaging for export. Other simple mechanisms that would explain why these diverse components may be brought in proximity with each other include nuclear ribosome biogenesis or retrograde tRNA transport. Finally, the coupling motif may suggest something beyond simple juxtaposition, presenting candidate proteins for the coupling of transcription to translation, and then from translation to mRNA export. This is strengthened further by the membership of the key mRNA export protein Yra1p, a protein previously implicated in coupling export to transcription, in the mediating cluster C70.

Our analyses suggest an extensive coupling of nuclear gene expression events, including the potential pioneer round of translation and NMD, tightly coupled to nuclear events from transcription to export through the nuclear pore. Coupling NMD to events that precede exit into the cytoplasm is a plausible possibility that presents significant benefits in efficiency, such as reducing the production of truncated polypeptides in the cytoplasm. The dedication of a nuclear fraction of translation and NMD machinery to quality control furthermore frees up the cytoplasmic fractions for mass protein production (Iborra *et al*, 2001). Furthermore, while the mammalian nuclear translation initiation factor 4A-like factor eIF4AIII is required for NMD, its cytoplasmic homolog eIF4A1/II is not—but is required for bulk translation (Ferraiuolo *et al*, 2004; Palacios *et al*, 2004; Shibuya *et al*, 2004). The yeast homolog of eIF4AIII is Fal1p, but to our knowledge its possible role in nuclear translation has not been addressed. Regardless, investigation of the mechanisms suggested by the motifs described above may offer further insight into the orchestration of translation and NMD.

Discussion

Eukaryotic gene expression, once viewed as a stepwise process involving distinct cellular machines, is now generally viewed as a series of highly coupled subprocesses (Ares and Proudfoot, 2005). While both biochemical and genetic experiments have identified functional interactions among a limited number of components of gene expression complexes, the immediate need for a systematic search for proteins involved in coupling has been recognized (Hieronymus and Silver, 2004). Our goal was to confirm and predict proteins coupling gene expression machines in yeast, and to extrapolate where possible to mammalian genomes. To accomplish this, we have created a modular and extensible framework to integrate large sets of data and generate a priority list of human-interpretable and readily testable hypotheses, each of which may be difficult or impossible to predict by manual inspection. We sought to develop a general method that is applicable to all organisms, although here we made use of yeast protein interaction data because they are by far the most extensively available databases. We recognize that the increased complexity of mammalian gene expression clearly involves coupling mechanisms that are distinct from those in yeast. Still, our analysis over the yeast data and the development of computational methods provide the framework for future analyses of coupling of gene expression in mammalian cells.

Comparison of protein interaction data set integration and network clustering methods and outcomes is an ongoing challenge (Gerstein *et al*, 2002; Hart *et al*, 2005). We took a principled approach to identifying potentially long-range coupling among distinct machines given noisy data. Significantly, two of the methods presented are novel contributions that should be applicable to other research applications: (1) our RDQ method of interaction data set quality calculation introduces a principled way to evaluate data sets independent of reference data and (2) the development of a pairwise CC for weighted graphs significantly improves upon existing ways to measure local neighborhood cohesiveness in an interaction network.

There are several limitations of our approach. First, we incorporate diverse data sets that include a broad spectrum of cell and experimental conditions, and therefore lack the ability to distinguish behaviors and protein interactions specific to various cell states such as meiosis, which involves degradation of the nuclear membrane. Second, the RDQ method presented here may be used to evaluate any number of data sets of similar type (i.e., protein interaction); however, if it were desired to extend the method to include dissimilar and uncorrelated data sets such as mRNA coexpression (Jansen *et al*, 2002), their RDQ values must be evaluated separately to avoid underweighting due to lack of data set overlap. Third, in our data representation, each protein is represented exactly once, a simplification that overlooks complicating factors such as copy number and appearance of each individual protein at various locations in the cell. Finally, our approach necessitates demarcation of discrete clusters, although the composition of protein complexes may be dynamic. Other approaches to data clustering (Samanta and Liang, 2003; Spirin and Mirny, 2003) may help address these challenges, and as methods for comparing clustering results are developed (Hart *et al*, 2005), this step in our approach may be improved.

As in all computational analyses, care must be taken in interpreting the results through a lens of informed biological skepticism. Coupling motifs identified in the static network model may be due to true coupling (spatiotemporal conjunction along with functional cooperation), but may also be due to simple spatial conjunction, or multiple interactions of possibly multiple copies of a single protein. For example, in the coupling predicted for translation or NMD, alternative explanations include coupling of cytoplasmic translation with transcription through co-export of factors with mRNPs—an example of conjunction without cooperation. In an example addressing multiple independent interactions, cluster #23 includes two distinct protein subsets: one nuclear and one cytoplasmic. These are likely related to the two spatially and functionally distinct roles of their common binding partner Sas10p: chromatin silencing (Kamakaka and Rine, 1998) and ribosomal processing (Dragon *et al*, 2002).

Our analysis includes binding partners of proteins annotated to the gene expression pathway. This allows the possibility of identifying functional roles or coupling links for proteins in non-nuclear cellular locations (for a review of surprising nuclear roles for cytoplasmic and membrane proteins, see Benmerah *et al*, 2003). However, top-ranking coupling motifs implicating these diverse proteins may also stem from spurious interaction data or ideosyncracies of the analysis. As a monomer, for example, the actin protein Act1p plays a role in gene expression as a catalytic part of the chromatin-modifying machineries INO80 (Shen *et al*, 2003), Swr1 (Mizuguchi *et al*, 2004), and NuA4 (Galarneau *et al*, 2000). In our analysis, the elongation and chromatin remodeling-annotated cluster #20 contains 49 proteins, which all bind actin according to at least one of the protein interaction data sets used in this study. The cluster additionally contains extranuclear proteins such as cofilin, twinfilin, and myosin, which interact with actin in the context of its ubiquitous role as a structural polymer. The assignment of these actin polymer-related proteins to cluster #20 and the significant coupling links to the actin protein that they define result from the disparate functions of actin in various contexts, a critical difference that cannot be captured by our model.

Despite these limitations, we were able to confirm known coupling in yeast, such as that of transcription and RNA processing with export. Our results also predicted highly significant connections among the processes of transcription, translation, mRNA export, and NMD, providing a computational blueprint for composition and organization of machinery in the gene expression pathway. In addition, we used new, independently generated biological data to verify both the formation of functional clusters in the network and the identification of significant links between them. Thus, the key product of this automated and objective coupling analysis is a set of high-confidence predictions that provide a prioritized agenda for experimental validation.

New data sets and annotations can be readily incorporated into our analysis to allow ongoing improvements in the precision and accuracy of the coupled cluster map and predictions derived from it. Recent studies of protein interaction networks indicate a high degree of conservation in network structures that are not apparent from other genome features (Sharan *et al*, 2005), suggesting that our results in yeast may be extrapolated to other organisms. Above all, our methods are general and may be applied directly to data from studies on other organisms, as well as to study cellular pathways and processes in addition to the gene expression pathway.

Materials and methods

RDQ calculation

We measured pairwise data set overlap by defining $\mathbf{M}(g, h)$ as the percentage of data set g covered by data set h, or zero where $g=h$. RDQ values are given by the entries of the vector $\mathbf{X}_R$, given by solving the equation $\mathbf{M}\mathbf{X}_R=\lambda_R\mathbf{X}_R$, where λ_R and $\mathbf{X}_R$ are the dominant eigenvalue and the corresponding (principal) eigenvector, respectively, of $\mathbf{M}$ (Supplementary information 5).

Pairwise CC

Our formula measures the sum total weights of edges to each other and to neighbors in common (to account for false negatives), normalized by the sum total weights of all edges containing either protein (to compensate for false positives) (Supplementary information 8).

k-means network clustering

k initial centroids were randomly chosen. Each node was assigned to the centroid to which it had the greatest link weight, and each centroid was chosen to maximize the cluster score. The total network score was defined as the sum of cluster scores

$$\sum_{i=1}^{k}\sum_{j=1}^{r_i} CC(c_i, n_{i_j})$$

where k is the number of clusters, r_i, c_i, and n_{i_j} are the size, centroid, and the jth member, respectively, of the ith cluster, and $CC(x,y)$ gives the pairwise CC between nodes x and y. Whenever a node has a CC score of zero with all of the current centroids, it is assigned to the centroid separated from it by the shortest path along links in the interaction network, determined using Dijkstra's algorithm (Cormen, 2001). Iterations of reassigning nodes to centroids and reassigning centroids to clusters were performed until convergence of the total network score. Convergence of total network score to within four digits was used in place of absolute convergence. The clustering process was repeated 70 times and the resulting clustering with highest total network score was selected.

Functional enrichment of clusters

The hypergeometric *P*-value (Tavazoie *et al*, 1999; Wu *et al*, 2002) was used to quantify the enrichment of each cluster in proteins annotated to categories corresponding to either the subprocesses of gene expression or GO-Slim categories. The *P*-value was calculated as follows:

$$P = 1 - \sum_{i=0}^{\min(k-1),(C-G+n)} \frac{\binom{C}{i}\binom{G-C}{n-1}}{\binom{G}{n}},$$

where G is the total number of proteins in the network, C the number of proteins in the cluster, n the number of proteins annotated to the tested category, and k the number of proteins in the cluster annotated to that subprocess. For each cluster, the expected value of the number of proteins in each category was calculated as well. When the actual number of proteins in a category was less than the expected value, the calculated *P*-value instead reflected the statistically significant lack of enrichment, and was replaced by $1-P$ to correct for this fact.

Motif finding

The cluster separability score was defined as

$$\frac{\sum_{r=1}^{Si} CC(c_i, n_{i_r}) + \sum_{r=1}^{s_j} CC(c_j, n_{i_j})}{\sum_{r=1}^{s_l} CC(c_l, n_{l_r})}$$

for a pair of clusters i and j, where s_i, c_i, and n_{i_r} are the size, centroid, and the rth member, respectively, of the ith cluster, and l is the cluster formed by merging clusters i and j and finding its new centroid.

Acknowledgements

We thank Natalie Thompson and Thanuja Premwardena for extraction of validation data, Naoko Tanese for comments, George Church, Aviv Regev, and June Oshiro for helpful discussion, and Guocheng Yuan, Nicole Francis, and Barbara Wold for critical reading of the manuscript. We also thank the anonymous reviewers at Nature MSB for valuable and constructive critiques. We thank the Bauer Center for Genomics Research for its generous support of LFW, SJA, and KM.

Contributions

KM, SJA, LFW, and TM designed the study and analyzed the results. KM and MS implemented the computational framework. NJK, AE, and JFG supplied the validation data. KM, SJA, LFW, and TM wrote the paper.

References

Aguilera A (2005) Cotranscriptional mRNP assembly: from the DNA to the nuclear pore. *Curr Opin Cell Biol* **17:** 242–250

Albertini M, Pemberton LF, Rosenblum JS, Blobel G (1998) A novel nuclear import pathway for the transcription factor TFIIS. *J Cell Biol* **143:** 1447–1455

Ares Jr M, Proudfoot NJ (2005) The spanish connection: transcription and mRNA processing get even closer. *Cell* **120:** 163–166

Auty R, Steen H, Myers LC, Persinger J, Bartholomew B, Gygi SP, Buratowski S (2004) Purification of active TFIID from *Saccharomyces cerevisiae*. Extensive promoter contacts and co-activator function. *J Biol Chem* **279:** 49973–49981

Bader GD, Hogue CW (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat Biotechnol* **20:** 991–997

Benmerah A, Scott M, Poupon V, Marullo S (2003) Nuclear functions for plasma membrane-associated proteins? *Traffic* **4:** 503–511, Review

Bentley DL (2005) Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Curr Opin Cell Biol* **17:** 251–256

Brogna S, Sato TA, Rosbash M (2002) Ribosome components are associated with sites of transcription. *Mol Cell* **10:** 93–104

Burckin T, Nagel R, Mandel-Gutfreund Y, Shiue L, Clark TA, Chong JL, Chang TH, Squazzo S, Hartzog G, Ares Jr M (2005) Exploring functional relationships between components of the gene expression machinery. *Nat Struct Mol Biol* **12:** 175–182

Colot HV, Stutz F, Rosbash M (1996) The yeast splicing factor Mud13p is a commitment complex component and corresponds to CBP20, the small subunit of the nuclear cap-binding complex. *Genes Dev* **10:** 1699–1708

Cormen TH (2001) *Introduction to Algorithms*. Cambridge, MA: MIT Press

Czaplinski K, Ruiz-Echevarria MJ, Paushkin SV, Han X, Weng Y, Perlick HA, Dietz HC, Ter-Avanesyan MD, Peltz SW (1998) The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs. *Genes Dev* **12:** 1665–1677

Dahlberg JE, Lund E (2004) Does protein synthesis occur in the nucleus? *Curr Opin Cell Biol* **16:** 335–338, Review

Darzacq X, Singer RH, Shav-Tal Y (2005) Dynamics of transcription and mRNA export. *Curr Opin Cell Biol* **17:** 332–339

de la Cruz J, Kressler D, Linder P (1999) Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families. *Trends Biochem Sci* **24:** 192–198

Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* **1:** 349–356

Dragon F, Gallagher JE, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, Beyer AL, Hunt DF, Baserga SJ (2002) A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* **417:** 967–970

Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* **18:** 529–536

Estruch F, Cole CN (2003) An early function during transcription for the yeast mRNA export factor Dbp5p/Rat8p suggested by its genetic and physical interactions with transcription factor IIH components. *Mol Biol Cell* **14:** 1664–1676

Ferraiuolo MA, Lee CS, Ler LW, Hsu JL, Costa-Mattioli M, Luo MJ, Reed R, Sonenberg N (2004) A nuclear translation-like factor eIF4AIII is recruited to the mRNA during splicing and functions in nonsense-mediated decay. *Proc Natl Acad Sci USA* **101:** 4118–4123

Galarneau L, Nourani A, Boudreault AA, Zhang Y, Heliot L (2000) Multiple links between the NuA4 histone acetyltransferase complex and epigenetic control of transcription. *Mol Cell* **5:** 927–937

Gao Q, Das B, Sherman F, Maquat LE (2005) Cap-binding protein 1-mediated and eukaryotic translation initiation factor 4E-mediated pioneer rounds of translation in yeast. *Proc Natl Acad Sci USA* **102:** 4258–4263

Gartenberg MR, Neumann FR, Laroche T, Blaszczyk M, Gasser SM (2004) Sir-mediated repression can occur independently of chromosomal and subnuclear contexts. *Cell* **119:** 955–967

Gatfield D, Le Hir H, Schmitt C, Braun IC, Kocher T, Wilm M, Izaurralde E (2001) The DExH/D box protein HEL/UAP56 is essential for mRNA nuclear export in *Drosophila*. *Curr Biol* **11:** 1716–1721

Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141–147

Gerstein M, Lan N, Jansen R (2002) Proteomics. Integrating interactomes. *Science* **295:** 284–287

Goldberg DS, Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* **100:** 4372–4376

Goyer C, Altmann M, Lee HS, Blanc A, Deshmukh M, Woolford Jr JL, Trachsel H, Sonenberg N (1993) TIF4631 and TIF4632: two yeast genes encoding the high-molecular-weight subunits of the cap-binding protein complex (eukaryotic initiation factor 4F) contain an RNA recognition motif-like sequence and carry out an essential function. *Mol Cell Biol* **13:** 4860–4874

Grigoriev A (2003) On the number of protein–protein interactions in the yeast proteome. *Nucleic Acids Res* **31:** 4157–4161

Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sonnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F (2005) Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436:** 861–865

Hammell CM, Gross S, Zenklusen D, Heath CV, Stutz F, Moore C, Cole CN (2002) Coupling of termination, 3′ processing, and mRNA export. *Mol Cell Biol* **22:** 6441–6457

Hart CE, Sharenbroich L, Bornstein BJ, Trout D, King B, Mjolsness E, Wold BJ (2005) A mathematical and computational framework for quantitative comparison and integration of large-scale gene expression data. *Nucleic Acids Res* **33:** 2580–2594

Hartigan JA (1975) *Clustering Algorithms*. New York: Wiley

He F, Jacobson A (1995) Identification of a novel component of the nonsense-mediated mRNA decay pathway by use of an interacting protein screen. *Genes Dev* **9:** 437–454

Hieronymus H, Silver PA (2004) A systems view of mRNP biology. *Genes Dev* **18:** 2845–2860

Hilleren P, Parker R (1999) Mechanisms of mRNA surveillance in eukaryotes. *Annu Rev Genet* **33:** 229–260

Hirose Y, Manley JL (2000) RNA polymerase II and the integration of nuclear events. *Genes Dev* **14:** 1415–1429

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415:** 180–183

Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. *Nature* **425:** 686–691

Iborra FJ, Escargueil AE, Kwek KY, Akoulitchev A, Cook PR (2004a) Molecular cross-talk between the transcription, translation, and nonsense-mediated decay machineries. *J Cell Sci* **117:** 899–906

Iborra FJ, Jackson DA, Cook PR (2001) Coupled transcription and translation within nuclei of mammalian cells. *Science* **293:** 1139–1142

Iborra FJ, Jackson DA, Cook PR (2004b) The case for nuclear translation. *J Cell Sci* **117:** 5713–5720

Ishigaki Y, Li X, Serin G, Maquat LE (2001) Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell* **106:** 607–617

Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* **7:** 535–545

Jansen R, Greenbaum D, Gerstein M (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res* **12:** 37–46

Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302:** 449–453

Jensen TH, Patricio K, McCarthy T, Rosbash M (2001) A block to mRNA nuclear export in *S. cerevisiae* leads to hyperadenylation of transcripts that accumulate at the site of transcription. *Mol Cell* **7:** 887–898

Kamakaka RT, Rine J (1998) Sir- and silencer-independent disruption of silencing in *Saccharomyces* by Sas10p. *Genetics* **149:** 903–914

Kim M, Ahn SH, Krogan NJ, Greenblatt JF, Buratowski S (2004) Transitions in RNA polymerase II elongation complexes at the 3′ ends of genes. *EMBO J* **23:** 354–364

Kim YJ, Bjorklund S, Li Y, Sayre MH, Kornberg RD (1994) A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* **77:** 599–608

Kuperwasser N, Brogna S, Dower K, Rosbash M (2004) Nonsense-mediated decay does not occur within the yeast nucleus. *RNA* **10:** 1907–1915

Kuras L, Borggrefe T, Kornberg RD (2003) Association of the Mediator complex with enhancers of active genes. *Proc Natl Acad Sci USA* **100:** 13887–13891

Lang V, Zanchin NI, Lunsdorf H, Tuite M, McCarthy JE (1994) Initiation factor eIF-4E of *Saccharomyces cerevisiae*. Distribution within the cell, binding to mRNA, and consequences of its overproduction. *J Biol Chem* **269:** 6117–6123

Laroche T, Martin SG, Tsai-Pflugfelder M, Gasser SM (2000) The dynamics of yeast telomeres and silencing proteins through the cell cycle. *J Struct Biol* **129:** 159–174

Le Hir H, Moore MJ, Maquat LE (2000) Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon–exon junctions. *Genes Dev* **14:** 1098–1108

Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. *Science* **306:** 1555–1558

Lei EP, Krebber H, Silver PA (2001) Messenger RNAs are recruited for nuclear export during transcription. *Genes Dev* **15:** 1771–1782

Lewis JD, Gorlich D, Mattaj IW (1996) A yeast cap binding protein complex (yCBC) acts at an early step in pre-mRNA splicing. *Nucleic Acids Res* **24:** 3332–3336

Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543

Liu Y, Guo W, Tartakoff PY, Tartakoff AM (1999) A Crm1p-independent nuclear export path for the mRNA-associated protein, Npl3p/Mtr13p. *Proc Natl Acad Sci USA* **96:** 6739–6744

Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506

Maquat LE (2002) Nonsense-mediated mRNA decay. *Curr Biol* **12:** R196–R197

Masuda S, Das R, Cheng H, Hurt E, Dorman N, Reed R (2005) Recruitment of the human TREX complex to mRNA during splicing. *Genes Dev* **19:** 1512–1517

Mendell JT, ap Rhys CM, Dietz HC (2002) Separable roles for rent1/hUpf1 in altered splicing and decay of nonsense transcripts. *Science* **298:** 419–422

Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30:** 31–34

Minvielle-Sebastia L, Keller W (1999) mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr Opin Cell Biol* **11:** 352–357

Misteli T (2001) Protein dynamics: implications for nuclear architecture and gene expression. *Science* **291:** 843–847

Mizuguchi G, Shen X, Landry J, Wu WH, Sen S, Wu C (2004) ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. *Science* **303:** 343–348

Nazarenus T, Cedarberg R, Bell R, Cheatle J, Forch A, Haifley A, Hou A, Wanja Kebaara B, Shields C, Stoysich K, Taylor R, Atkin AL (2005) Upf1p, a highly conserved protein required for nonsense-mediated mRNAdecay, interacts with the nuclear pore proteins Nup100p and Nup116p. *Gene* **345:** 199–212

Orphanides G, Reinberg D (2002) A unified theory of gene expression. *Cell* **108:** 439–451

Page L, Brin S, Motwani R, Winograd T (1998) *The Page-Rank Citation Ranking: Bringing Order to the Web*, Stanford Digital Libraries Working Paper

Palacios IM, Gatfield D, St Johnston D, Izaurralde E (2004) An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay. *Nature* **427:** 753–757

Piruat JI, Aguilera A (1998) A novel yeast gene, THO2, is involved in RNA pol II transcription and provides new evidence for transcriptional elongation-associated recombination. *EMBO J* **17:** 4859–4872

Proudfoot NJ, Furger A, Dye MJ (2002) Integrating mRNA processing with transcription. *Cell* **108:** 501–512

Qi Y, Klein-Seetharaman J, Bar-Joseph Z (2005) Random forest similarity for protein–protein interaction prediction from multiple sources. *Pac Symp Biocomput* **10:** 531–542

Reed R (2003) Coupling transcription, splicing and mRNA export. *Curr Opin Cell Biol* **15:** 326–331

Reed R, Cheng H (2005) TREX, SR proteins and export of mRNA. *Curr Opin Cell Biol* **17:** 269–273

Reichert VL, Le Hir H, Jurica MS, Moore MJ (2002) 5′ exon interactions within the human spliceosome establish a framework for exon junction complex structure and assembly. *Genes Dev* **16:** 2778–2791

Robinson MM, Yatherajam G, Ranallo RT, Bric A, Paule MR, Stargell LA (2005) Mapping and functional characterization of the TAF11 interaction with TFIIA. *Mol Cell Biol* **25:** 945–957

Rodriguez-Navarro S, Fischer T, Luo MJ, Antunez O, Brettschneider S, Lechner J, Perez-Ortin JE, Reed R, Hurt E (2004) Sus1, a functional component of the SAGA histone acetylase complex and the nuclear pore-associated mRNA export machinery. *Cell* **116:** 75–86

Ruiz-Echevarria MJ, Gonzalez CI, Peltz SW (1998) Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA. *EMBO J* **17:** 575–589

Samanta MP, Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci USA* **100:** 12579–12583

Schmitt C, von Kobbe C, Bachi A, Pante N, Rodrigues JP, Boscheron C, Rigaut G, Wilm M, Seraphin B, Carmo-Fonseca M, Izaurralde E (1999) Dbp5, a DEAD-box protein required for mRNA export, is recruited to the cytoplasmic fibrils of nuclear pore complex via a conserved interaction with CAN/Nup159p. *EMBO J* **18:** 4332–4347

Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) From the Cover: conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* **102:** 1974–1979

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31:** 64–68

Shen X, Ranallo R, Choi E, Wu C (2003) Involvement of actin-related proteins in ATP-dependent chromatin remodeling. *Mol Cell* **12:** 147–155

Shibuya T, Tange TO, Sonenberg N, Moore MJ (2004) eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense-mediated decay. *Nat Struct Mol Biol* **11:** 346–351

Sommer P, Nehrbass U (2005) Quality control of messenger ribonucleoprotein particles in the nucleus and at the pore. *Curr Opin Cell Biol* **17:** 294–301

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* **100:** 12123–12128

Strasser K, Masuda S, Mason P, Pfannstiel J, Oppizzi M, Rodriguez-Navarro S, Rondon AG, Aguilera A, Struhl K, Reed R, Hurt E (2002) TREX is a conserved complex coupling transcription with messenger RNA export. *Nature* **417:** 304–308

Tanay A, Sharan R, Kupiec M, Shamir R (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* **101:** 2981–2986

Tange TO, Nott A, Moore MJ (2004) The ever-increasing complexities of the exon junction complex. *Curr Opin Cell Biol* **16:** 279–284

Tarun Jr SZ, Wells SE, Deardorff JA, Sachs AB (1997) Translation initiation factor eIF4G mediates *in vitro* poly(A) tail-dependent translation. *Proc Natl Acad Sci USA* **94:** 9046–9051

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* **22:** 281–285

Tornow S, Mewes HW (2003) Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res* **31:** 6283–6289

Vazquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabasi AL (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci USA* **101:** 17940–17945

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417:** 399–403

Walker SS, Shen WC, Reese JC, Apone LM, Green MR (1997) Yeast TAF(II)145 required for transcription of G1/S cyclin genes and regulated by the cellular growth state. *Cell* **90:** 607–614

Weirich CS, Erzberger JP, Berger JM, Weis K (2004) The N-terminal domain of Nup159 forms a beta-propeller that functions in mRNA export by tethering the helicase Dbp5 to the nuclear pore. *Mol Cell* **16:** 749–760

Wilkinson MF, Shyu AB (2002) RNA surveillance by nuclear scanning? *Nat Cell Biol* **4:** E144–E147

Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat Genet* **31:** 255–265

Yeger-Lotem E, Margalit H (2003) Detection of regulatory circuits by integrating the cellular networks of protein–protein interactions and transcription regulation. *Nucleic Acids Res* **31:** 6053–6061